ISSN: 0976-3104
SPECIAL ISSUE: (EMERGING TECHNOLOGIES IN NETWORKING AND SECURITY (ETNS)

ARTICLE OPEN ACCESS

AN INTELLIGENT FEATURE SELECTION APPROACH FOR GENE EXPRESSION DATA USING HYBRID BIOGEOGRAPHY BASED OPTIMIZATION (BBO) WITH ARTIFICIAL BEE COLONY (ABC) ALGORITHM

S Venkata Krishna Kumar[1*] and R Nedunchezhian[2]
[1]Dept. of Computer Science, PSG College of Arts and Science, Coimbatore, Tamil Nadu, INDIA
[2] Dept.of Computer Science and Engineering, Kalaingar Karunanidhi Institute of Technology, INDIA

ABSTRACT

Micro array data's plays major role in the bio medical fields which is used to analyze and predict the various diseases effectively. Classification is a most concerned data mining strategy that can be used to predict the disease by classifying the micro array data in terms of their feature values. However, this would be more difficult task due to presence of irrelevant information present in the micro array data such as noises, redundant genes, and uninformative genes and so on. This research work focus on implementing the methodology which can be used to classify the micro array data accurately with increased performance result. Methods:In the proposed research methodology optimal feature selection is focused to avoid the irrelevant information given as input to the classifier, thus the performance can be optimized. It is achieved by introducing the novel mechanism namely Hybridized biogeography based optimization (BBO) and Artificial bee colony algorithm (ABC) in which migration operator process is combined with the ABC algorithm. In the proposed research method, preprocessing is done by using NLLS impute algorithm which would select the top most 60 genes from the set of gene with replacement of missing values. After pre-processing SVM-REF and BBF strategies are used for the classification and gene selection purpose that is finished with the assistance of leave-one-out cross validation methodology (LOOCV). Results:The evaluation of the proposed research methodology is done in the matlab simulation environment under varying size of micro array data set. The results are compared against the different performance metrics with various existing research works for gene expression dataset bench marks. Conclusions:The finding of the research work proves that proposed method outperforms the existing methodologies.Applications: This approach would be more useful in the fields such as bioscience, bioengineering,and medical fields for the appropriate detection and analysis of the various diseases and their patterns.

Published on: 2nd -December-2016

KEY WORDS

Gene expression data, gene selection, hybrid approach, multi-objective binary biogeography based optimization

***Corresponding author:** **Email:** leenasilvoster@gmail.com **Tel.:** +91-7385644304; **Fax:** 020 24351308

INTRODUCTION

Gene selection is one of the most important technique to find the informative genes from the cells. However it is most difficult to predict the genes from the large volume of data's that are irrelevant to the cancer disease. This can be resolved by integrating the gene selection approach which is used to find the informative genes and omit the irrelevant gene data's [1].

The gene expression data like non-relevant data, noise data and comparison of gene samples (high dimensional data) causes tedious problem in the gene selection process [2]. To avoid this problem, the efficient classifier's are required to classify the exact samples of genes in a particular set of instructive genes from the part of a larger set of gene samples by using a gene selection method. The gene selection is called feature selection in the computational intelligence domain [3]. The advantages of gene selection are

1) It improves the accuracy of the classification,
2) Shorter process time.
3) It can reduce the over fitting of the data.
4) It can take away the not related and noise genes.

The feature selection method is mainly used to decrease the more number of genes in classification in the classification accuracy [4,5]. It means this method mainly involved increasing the classification performance and decreasing the more number of features into small number of features selected. Generally the feature selection based on single objective problem based evolutionary algorithms is not as much of common and computationally costly as multi objective approached.For that reason the evolutionary algorithm is used to develop a multi objective approach is required and at the same time it make the most of the classification rate and reduce the number of features selected.

The NSGA-II [6] and SPEA [7] are currently planned for the feature selection in multi objective based methods. To get small set of non-redundant disease related genes by using the multi objective particles swarm optimization from the hybrid multi-objective optimization method [8]. Salcedo-Sanz et al uses multi objective genetic algorithms used to common uniqueness of the samples as feature correlation and search the subset of features by combining different filter approaches criteria for feature selection [9]. The multi-objective approach in hybrid GA and support vector machine classifier (GASVM) is proposed for gene selection and the classification of gene expression data [10]. An archived multi-objective simulated annealing (AMOSA) is developed for predicting miRNA promoters using an SVM with RBF kernel in a feature selection method [11, 12].

A neural network is working to let the use of a representative database to calculatesuitability.A multi-objective evolutionary algorithm is projected to solve the difficulty of gene selection in the gene subset size minimization and performance maximization [13]. To establish the benchmark problems by using the optimization of split modified radius-margin model selection criteria [14].The a multi-objective genetic algorithm is used to gene selection in the microarray datasets [15]. This process is consummate by support vector machines. A multi objective genetic algorithm used to increase the classification accuracy rate in the number of features [16]. And also the multi objective genetic algorithm useful for image processing especially for edge detection and the complete comparison on the canny edge detector [17]. In the bioinformatics and computational biology field using multi objective optimization technique [18]. This technique is mainly used to elucidate the reasons after the use of multi objective optimization in each application area and also to direct the possible futures.

MATERIALS AND METHODS

In the proposed research methdology, an improved classifier for the micro array data to diagnize the disease is introduced which is achieved by introducing the optimal feature selection mechanism that can avoid the irrelevant features from the set of micro array data set.In this approach gene expression data set is given as a input. Initially infomax ICA mechanism is used to select the top 60genes from the input data. Then the preprocessing is done to replace the missing values using NLLS imputation approach. Then the SVM-REF and BBF is used to perform gene selection by adapting leave one at last out cross validation (LOOCV) method in the support vector machine. At last HABCBBO is adapted to perform optimal feature selection with the concern of discrete problems. It is called BBHK and is based on migration model. The overall proposed methodology shown in the [Figure-1].

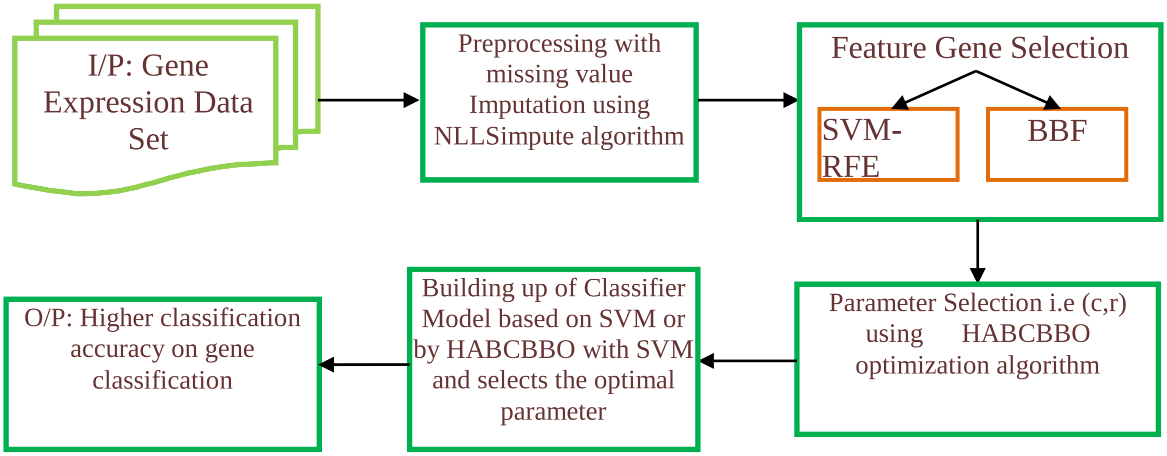
I/P: Gene Expression Data Set
Preprocessing with missing value Imputation using NLLSimpute algorithm
Feature Gene Selection
SVM-RFE
BBF
Parameter Selection i.e (c,r) using HABCBBO optimization algorithm
Building up of Classifier Model based on SVM or by HABCBBO with SVM and selects the optimal parameter
O/P: Higher classification accuracy on gene classification

Fig.1: Overall Proposed Methodology

Preprocessing of the dataset with missing value imputation

The gene expression microarray data is preprocessed to reduce the larger amount of genes to the small set of genes. It is used to increase the competnce of classification accuracy rate in the high dimenional data set. For various tentative reasons the microarray data might contains a missing values. Before applying the microarray data in data analysis algorithm it is very important to implicate the missing values. In this proposed work a New Local Least Square imputatiom based algorithm (NLLSimpute) implemented.

THE IIOAB JOURNAL ISSN: 0976-3104

This algorithm is similar as K-nearest neighbor imputatiom algorithm.In this research work it is used to select the K-nearest gene from the complete set of obtained candidate gene as a substitute of the complete genes and also reuse the imputed information available. Subsequently the highest missing rate will be imputed in the target gene with the available information. The New Local Least Square impute algorithm proceeds as follows:

```
Algorithm: NSLLSimpute

    Input: The m number of ganes and n number of samples from the  microarray gene expression matrix T it  contains some artificial missing entries.
    Output: The missing entries are predicted.
    1) At first from the complete gene expression matrix X to replace the missing positions of the given gene expression matrix.
     2) According to their missing rate and imputation one after another the m number of genes in X are sorted in ascending order.
     3) For each target gene gt in X do:
    a) For each missing position p in target gene gt do:
    i) The Pearson Correlation coefficient is used to select the k number of candidate genes nearest to the target gene.for the period of the computation of Pearson Correlation coeffient the missing positions filled with row average are ignored[9].
    ii) The conception of local least squares is imputing the missing position p in the target gene beneathdeliberation[9].
     iii) The result of imputed values are positioned in the missing positions of the considered target gene.
     4) End
```

Gene selection

The integrated SVM-RFE having a excellent recital on classification but it having the poor performance on seperability in redundant class labels. BBF is used to avoids redundant class labels in selection. Both combined achieves as good asrecital. The Fig.1 shows the schematic view of overall process carried. The SVM-RFE and BBF are used to select the feature from dataset after selection the features are going to the process of classifier for training. Finally in the training session the evaluation conceded with testing data.

Feature Selection Method

The recursive elimination procedure of SVM-RFE [3] is implemented as follows:

```
1. Start: ranked set R=[ ]; picked feature subset S=[1, …, d].
2. Repeat until all features in subset gets ranked:
a. Train the features with SVM from set S as input variables.
b. Calculate the weight vector for each feature.
c. Calculate the ranking score for features in set S.
d. Identify the feature with the smallest ranking score.
e. Update.
f. Eliminate smallest ranking feature.
3. Result: Ranked feature set R[ ].
```

After ranking genes by SVM-RFE we eliminate redundancy by applying Based Bayes Error Filter (BBF). The BBF intially processed the relevant candidate genes that are selected by a criterion function and second the criterion controlling the upper bound of the Bayes error is applied to the relevant candidate genes in order to remove the redundant genes. By this way, the genes and the subset of the gene are selected for further processing.

HABCBBO for optimizing the objective function

The stages of ABC algorithms are categorized into three phases. Those are empploye bee phase, onlooker bee phase, scout bee phase. Initiallly, The population of bee N p would be created randomly by using the bees that are present in the search space Ω. This population is nothing but the group of food sources. Based on the volume of the food space, bees attraction would differ. For example, food space with volume would be attracted more by the bees and less food volume would be less attracted by bees. Thus in ABC appraoch, objective function is find an food source with more volume of food. The fitness funciton based on this objective fucntion is calculated after initializaiton of the bee population. The equation used to calculated the fitness fucntion is givne as follows:

$$\mathrm{fit}(X_i) = \begin{cases} \dfrac{1}{1+f(X_i)} \mathrm{iff}(X_i) \geq 0 \\ 1 + |f(X_i)| \mathrm{Otherwise} \end{cases}$$

In employee phase, swarm of employess bees would be sent to find an food source that contains more volume of food. In this swarm of employees bees, atleast one employee bee would find the food source where the bee that find the food is denoted as i and the food source that was found is vi.After finding the new food source with more volume of food, the old food source position would be updated by substituting the new xi in old value.Onlooker bees are used to filter the food sources that are found by the employee bees based on the food quality and the volume. This is calculated by using the roulette wheel selection approach which finds an food source based on probability values. Scout bees are used in the emergency situation where the optimal soluiton cannot be found. It will send out to find an food source when the employee and onlooker bees cannot find an opitmal solution for the particular period of time [10, 11, 12]. It is employed based on the parameter called ‘limit’ in terms of time unit. When the optimal food source cannot be obtained in the limit time period then the scout bees would be employed.The ABC algorithm lacks in solution updation where the solution are not updated in the intermediate level like cross over, mutation. Thus the solution identification would not produce optimal solution [16]. This can be improved by integrating the biogeography-based optimization (BBO) approach for the migration of solution in the intermediate level.

Information migration scheme

The BBO scheme is based on the biological behaviour of the sea species isolation from the one locaiton to another location. The measure that is used in the BBO process for finding the optimal solution is habitat suitability index (HSI). Here the solution that have more HSI is defined as good solution, less HSI would be taken as poor solution. In BBO approach, habitate with more emigration rate and the less immigration rate would be taken as best solution. Based on these HSI values, solutions would be sorted out and the calculation procedure of μ and λ is calculated using the following equation.

$$\lambda_i = I(1 - \frac{i}{n})$$

$$\mu_i = E\left(\frac{i}{n}\right)$$

Hybrid ABC with migration operator

As mentioned in the introduction, ABC does not do well in exploiting the existing information of solutions. On the other hand, BBO is good at exploiting the information of existing solutions. Based on these observations, we propose to hybridize ABC with migration operator of the BBO algorithm to combine their strengths, called HABCBBO. This algorithm combines the migration operator with the employed bee phase of ABC. In addition, Rechenberg's one-fifth success rule is employed to effectively control the adaptation of immigration probability.

Algorithm 1

Input: f(.), D, x^{min}, x^{max}, N_p, limit

Output: the best solution obtained from the algorithm

```
Randomly create $N_p$ solution $x_1, x_2, x_3, \dots \dots \dots \dots \dots, x_{N_p}$
Evaluate function values of solution and their fitness by (1).
Set counter $l_i$, i=1,2,.........$N_p$ to 0;
Repeat
    Send out employed bees (see algorithm 2);
    Send out onlooker bees depending on their nectar
    Amounts by (2);
    Evaluate candidate solution and their fitness by (1).
    Do greedy selection by (3) and update $l_i$ by (5)
    Send out scout bees if $l_i$ reaches limits
    Evaluate candidate solution and their fitness by (1).
    Reset $l_i$ to 0.
Until termination criterion are met
```

Algorithm 2: Pseudo code of the employed bee phase in HABCBBO

```
Set $S_c = 0$, $\sigma = 0.5\ D$ and $c_d = 0.82$
Calculate $\lambda_i$ and $\mu_i$ for each solution in the bee colony
For i=1 to $N_p$ do
    Randomly choose $j_1 \in [1, D]$ and $r_1 \in [1, N_p]$
    For j=1 to D do
        If j==$j_1$then
            $v_{i,j1} = x_{i,j1} + \varphi(x_{i,j1} - x_{r1,j1})$
        Else if rand (0,1)<$\frac{\lambda_i}{\sigma}$ then
            Choose a solution $x_{r2}$with $r2 \neq i$ using roulette wheel selection method based on emigration rates $\mu_i, i = 1,2,\dots, N_p$
            $v_{i,j} = x_{r2,j}$
        else
            $v_{i,j} = x_{ij}$
        else
    else for
    evaluate candidate solution and their fitness by (1)
    if $f(v_i) < f(x_i)$ then
        replace $x_i$ by $v_i$
```

```
$S_c + +$
End if
End for
If $\frac{S_c}{N_p} < \frac{1}{5}$ then
$\sigma = c_d . \sigma$
Elase If $\frac{S_c}{N_p} > \frac{1}{5}$ then
$\sigma = \frac{\sigma}{c_d}$
Else if
Reset $S_c = 0$
```

HABCBBO is described in Algorithm 1. This algorithm is same as ABC algorithm in which procedure differs in the phase of employee. Initially immigration and the emigration rate would be calculated by using equations qw and 13. And the optimal solution updation procedure is represented in the following equation.

$$v_{ij} = \begin{cases} x_{i,j} + \varphi\left(x_{i,j} - x_{r1,j}\right) \text{if} j = j_1 \\ x_{r2,j} \text{ifrand}(0,1) < \frac{\lambda_i}{\sigma} \text{and} j \neq j_1 \\ x_{i,j} \text{Otherwise} \end{cases}$$

The above procedure is repeated for the Np times to obtain the optimal solution. And the more opt best solution is found by comparing the best population with the location of the populations. If $f(V_i) < f(X_i)$ then the solution is said more optimal, else optimal solution is not obtained. During this iteration, number of successful operation would be counted and it is incremented by Sc and the one fifth rule is applied.

$$\sigma = \begin{cases} c_d . \sigma \text{if} \frac{S_c}{N_p} < \frac{1}{5} \\ \frac{\sigma}{c_d} \text{if} \frac{S_c}{N_p} > \frac{1}{5} \\ \sigma \text{Otherwise} \end{cases}$$

Where c_d → decay factor which is assumed as 0.82 for the adaptation[18].

In the hybridized employed bee phase, σ is used to control the adaptation of immigration rate. In Algorithm 2, σ is initialized to 0.5D, which is determined based on experiment on toy functions.The final result of the optimization is the best individual of the last iteration.

HABCBBO and support vector machines

The issues that are found in the SVM methodology areoptimal selection of input feature subset and the kernel parameters. This issue is resolved in the proposed research method by integrating the HABCBBO algorithm with the SVM approach which is called as HABCBBO-SVM to optimize the parameters C and r. By doing so, feature subset selection can be done optimally and the testing accuracy of SVM can be improved. In the first phase, the BBKH algorithm provides a binary encoded individual where each bit represents a gene. If a bit is 1, it denotes this gene is kept in the subset; else if at bit is 0, it represents a non-selected feature. Therefore, the individual length is equal to 2+D in the initial microarray dataset. Then, the fitness of each individual is assessed by the accuracy of leave-one-out cross-validation method (LOOCV). The leave-one-out cross-validation method can be described as follows: when there are n data to be classified, the data are divided into one testing sample and n-1 training samples. Each individual will be selected as a testing sample in turn. The other n-1 individuals serve as the training data set to determine the prediction parameter of the model. The proposed research methodology HABCBBO_SVM is demonstrated by using the data set that contains 7 records namely A1, A2, A3, A4, A5, A6, and A7 that contains four features.Among these records, six records are used for training process and the remaining one record is taken for testing process. The SVM method fixes the parameter values as $C = 2^5$ and$r = 2^{-2}$, thus the more classification accuracy can be obtained. The fitness calculation is calculated as like follows:

$$f_1 = \text{SVM\_accuracy}$$

$$f_2 = \left(\frac{D - R}{D}\right)$$

$$f = [f_1, f_2]$$

where SVM_accuracy→SVM classification accuracy, D → total number of the genes, and R → number of selected genes.

The process of the algorithm is as follows

Step 1: The gene expression data are pre-processed by the Infomax ICA method. The 60 top genes with the highest scores are selected as the crude gene subset. The corresponding subsets in the testing parts are also selected at the same time.

Step 2: Converting genotype to phenotype. This step will convert parameter C, r and feature subset from its genotype to a phenotype.

Step 3: the two objectives functions, f_1 and $f_2$in (9) are calculated using ABCBBO optimization.

Step 4: Multi-objective binary biogeography based optimization. In this step, the algorithm searches for better solutions by binary migration model and binary mutation model.

Step 5: Checking the termination criterion. The final feature subsets are selected, and then output the feature subset and the parameters C and r.

THE IIOAB JOURNAL ISSN: 0976-3104

The System architecture of the proposed HABCBBO based feature selection and parameters optimization for support vector machine is shown in **[Figure- 2].**

RESULTS

The relevant gene selection is the most complex task in the gene expression classification method. In this research work, ten gene expression data is selected for the classification that includes micro array data of tumor samples, brain tumor, leukemia, lung cancer, and prostate tumor samples. This misro array data can be downloaded from the website http://www.gems-sytem.org. The data sets are partitioned into one testing sample and n-1 training samples. Note that each individual will be selected as a testing sample in turn. The remaining n-1 individuals will then serve as the training data set.

Table: 1. Format of gene expression classification data

Dataset Number	Dataset Name	Number of Samples	Number of Genes	Number of Classes
1	11_Tumors	60	9	5726
2	9_Tumors	174	11	12533
3	Brain_Tumors 1	90	5	5920
4	Brain_Tumors 2	50	4	10367
5	Leukemia 1	72	3	5327
6	Leukemia 2	72	3	11225
7	Lung_Cancer	203	5	12600
8	SRBCT	83	4	2308
9	Prostate_Tumor	102	2	10509
10	DLBCL	77	2	5469

The proposed research work namely HABCBBO-SVM is implemented in the Matlab-7.9 language development environment which is extended using the libsvm that was originated by Chang and Li [28]. The computing processor characteristics are Pentium 3.0 GHz Processor with 1.0 GB of memory. The proposed research work is evaluated and compared with the existing research works namely SVM Grid search, IBPSO, HPSOTS, PSO/GA [29], and a multi-objective algorithm NSGA-II [6]. The parameter values that are set are, “Population size:50, Number fo generation:100, HIS:1, Mutation rate:0.5”.

To make the experiments more accurate, each algorithm will be run ten times for each gene data. After that, an average result of the ten independent runs is obtained and compared. As stated earlier, the main objective of the proposed research work is increasing classification accuracy and the number of genes selected. Thus in this multi objective problem, the final solution is taken by using the pareto front method in which any solution is not dominated by any other solutions. Pareto front optimal procedure is used to select the most optimal solution among the various number of solutions that can lead to high classification accuracy.

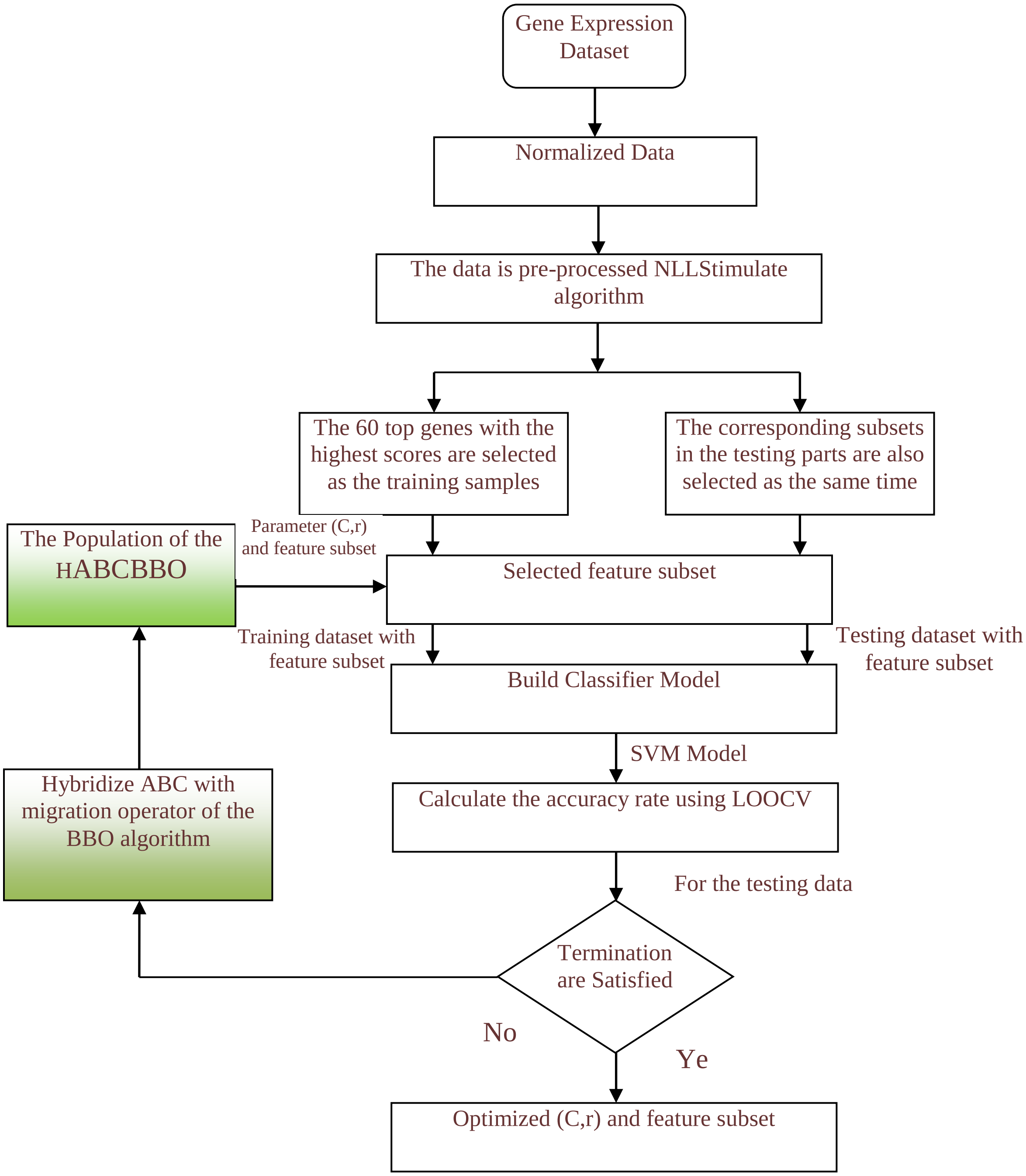
Gene Expression Dataset
Normalized Data
The data is pre-processed NLLStimulate algorithm
The 60 top genes with the highest scores are selected as the training samples
The corresponding subsets in the testing parts are also selected as the same time
The Population of the HABCBBO
Parameter (C,r) and feature subset
Selected feature subset
Training dataset with feature subset
Testing dataset with feature subset
Build Classifier Model
SVM Model
Hybridize ABC with migration operator of the BBO algorithm
Calculate the accuracy rate using LOOCV
For the testing data
Termination are Satisfied
No
Ye
Optimized (C,r) and feature subset

Fig. 2: System architectures of the proposed HABCBBO -based feature selection and parameters optimization for support vector machine

……………………………………………………………………………………………………

[Figure- 3, 4] represents the experiment results of MOBBKH SVM on ten gene datasets. In these figures, #acc denotes the testing accuracy, and #selected gene denotes the number of genes selected for these gene expression data. From the figures, we can see that the results of the proposed algorithm are consistent on all datasets. For the Leukemia1, Leukemia2, SRBCT, and DLBCL datasets, MOBBBO, the proposed MOBBKH can achieved 100% LOOCV accuracy with less than 10 selected genes. For the Brain_Tumor2 dataset, MOBBBO can obtainthe 100% LOOCV accuracy for nine times. For the other datasets of Lung Cancer, Prostate_Tumor, and Brain_Tumor1, the BBKF algorithm can also provide more than 96% classification accuracies except for the 11_Tumors dataset (92.414%) and 9_Tumors dataset (80.5%).

ISSN: 0976-3104

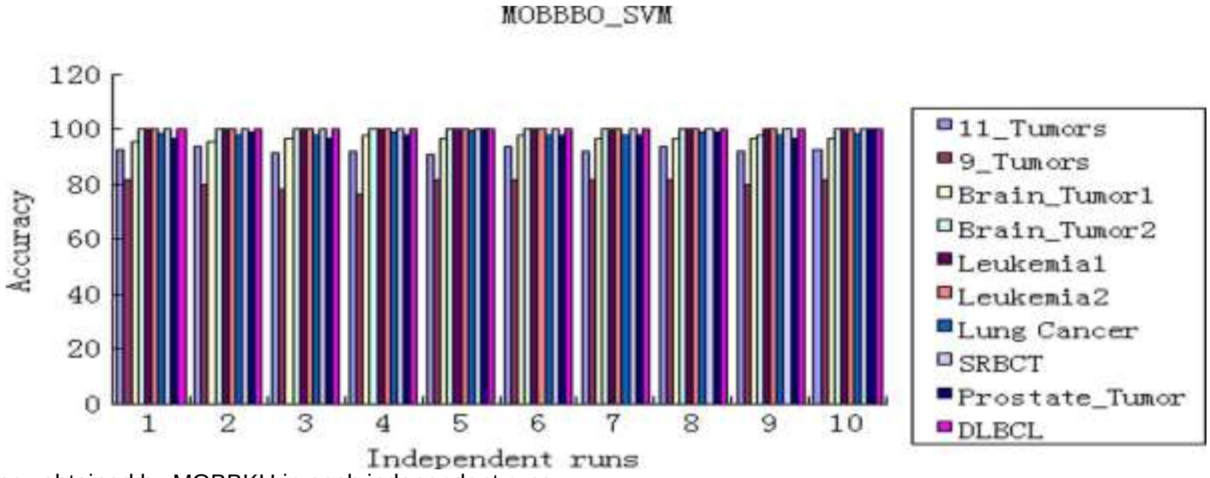

Fig. 3: The accuracy obtained by MOBBKH in each independent runs

………

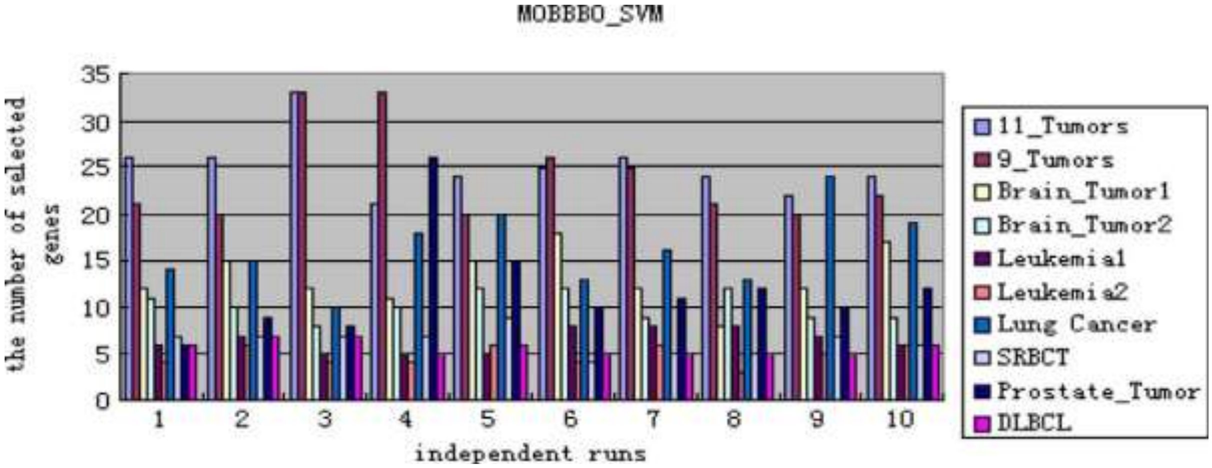

Fig. 4: The number of gene selected obtained by MOBBKH in each independent runs

………

The characteristics of the genes, selected genes, and percentage of genes selected percentage are shown in Table II. The average percentage of genes selected is 0.0017. For the Leukemia1, Leukemia2, SRBCT, and DLBCL dataset, our algorithm can obtain 100% LOOCV accuracy even though the percentage of genes selected is reduce to 0.0012, 0.0004, 0.0028, and 0.0010. So we can conclude that not all features are necessary for achieving the better classification accuracy.

Table 2:The genes, selected genes, and percentage of gene selected percentage

Dataset Name	Total Number f Genes	Number of Genes Selected	Percentage of genes selected
11_Tumors	5726	25.1	0.0044
9_Tumors	12533	24.1	0.0019
Brain_Tumors 1	5920	13.2	0.0022
Brain_Tumors 2	10367	10.2	0.0010
Leukemia 1	5327	6.5	0.0012
Leukemia 2	11225	4.8	0.0004
Lung_Cancer	12600	16.2	0.0013
SRBCT	2308	6.4	0.0028
Prostate_Tumor	10509	11.9	0.0011

In order to show the effective of each part, two experiments are conducted. The first experiment is to show the effective of the Fisher-Markov selector, and the second experiment is to show the effective of MOBBKH. In[Table 3], the better results between two algorithms are highlighted using boldface. It is easy to see that both the classification accuracy and the number of selected genes of HABCBBO SVM are superior to grid search SVM. This also demonstrates the effective of HABCBBO.

Table 3: Compared With mobbbo SVM With Other Previous PSO

Dataset Name	Evaluation	Proposed HABCBBO	MOBBKH	MOBBBO	IBPSO [8]	IBPSO [6]
11_Tumors	Accuracy	96	95	92.414	95.06	93.10
	Genes	26	26	25	240.9	2948
9_Tumors	Accuracy	82.5	81.23	80.5	75.50	78.33
	Genes	25	25	24.1	240.6	1280
Brain_Tumors 1	Accuracy	98	97.12	96.667	92.56	94.44
	Genes	14	14	13.2	11.2	754
Brain_Tumors 2	Accuracy	99.94	99.92	99.8	92.00	94.00
	Genes	11	11	10.2	9.10	1197
Leukemia 1	Accuracy	100	100	100	100	100
	Genes	7	7	6.5	3.5	1034
Leukemia 2	Accuracy	100	100	100	100	100
	Genes	5.2	5.2	4.8	6.7	1292
Lung_Cancer	Accuracy	99.5	99.1	98.473	95.86	96.55
	Genes	17	17	16.2	14.90	1897
SRBCT	Accuracy	100	100	100	100	100
	Genes	7.5	7.5	6.4	17.50	431
Prostate_Tumor	Accuracy	99.2	99	98.33	97.94	92.16
	Genes	13	13	11.9	13.60	1294
DLBCL	Accuracy	100	100	100	100	100
	Genes	6	6	5.7	6	1042

From the [Table-3] we can see for Leukemia1, Leukemia2, SRBCT, and DLBCL, both MOBBBO SVM and MOBBBO SVM without Fisher-Markov selector can obtain the 100% classification accuracy, while MOBBBO SVM can provide lower gene numbers. For 9_Tumors, MOBBBO SVM, and MOBBBO SVM without Fisher-Markov selector can obtain the 80.5% classification accuracy, and MOBBBO SVM can also provide lower gene number. For Brain_Tumors1, Brain_Tumors2, Lung_cancer, and Prostate_Tumor datasets, MOBBBO SVM can not only provide better classification accuracy, but also lower gene numbers. Only for the 11_Tumors, the MOBBBO without Fisher-Markov selector can generate the better classification accuracy, which also demonstrates that the Fisher-Markov selector is not suitable for all the situations. For the second experiment, we compare our approach with grid search SVM.

Based on the above analysis, the experimental results can demonstrate the flexibility and robustness of the proposed HABCBBO in feature selection. This algorithm can provide positive results when applied to the gene expression data with the limited number of features and samples. The reason may be that the multi-objective binary biogeography based optimization tends to share their features with low HSI solutions, which can accept a lot of new features from high HSI solutions. In HABCBBO, a habitat is a vector which follows binary migration and binary mutation step to the optimal solution. The new candidate habitat is generated from all the solutions in population by using the binary migration and binary mutation model. Following these rules, the HABCBBO algorithm finally produces a better subset for the gene classification.

CONCLUSION

In this paper, a hybrid multi-objective binary biogeography based optimization with support vector machine is proposed for gene selection on ten gene expression datasets. Experimental results show that the algorithm can simplify feature selection by finding a smaller number of features needed effectively and a higher classification accuracy compared with other previous methods. The proposed algorithm can obtain the highest accuracy in nine of the ten microarray dataset problems since the multi-objective approach in it can find a diverse solution inPareto optimal set. Moreover, the results show that there are many irrelevant genes in gene expression data and some of them are not relevant to a given cancer. For further work, the proposed algorithm can be applied to some problems in other fields.

CONFLICT OF INTEREST

Authors declare no conflict of interest

ACKNOWLEDGEMENT

None

FINANCIAL DISCLOSURE

No financial support was received to carry out this project

REFERENCES

[1] Liang Q, Cheng X, Samn S.[2010] NEW: Network-enabled electronic warfare for target recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 46(2):558–568.

[2] Liang Q, Cheng X, Huang S, Chen D. [2014] Opportunistic sensing in wireless sensor networks: theory and application. Comput. IEEE Trans.63(8):2002–2010.

[3] Singh S, Liang Q, Chen D, Sheng L.[2011] Sense through wall human detection using UWB radar, *Journal on Wireless Communications and Networking* 2011(1):503–520.

[4] Liang Q.[2011] Situation understanding based on heterogeneous sensor networks and human-inspired favor weak fuzzy logic system. *IEEE Systems Journal*. 5(2):156–163.

[5] Liang Q.[2011] Radar sensor wireless channel modeling in foliage environment: UWB versus narrowband. IEEE Sensors Journal 11(6):1448–1457.

[6] Xu L, Liang Q.[2012] Zero correlation zone sequence pair sets for MIMO radar. Aerospace and Electronic Systems, *IEEE Transaction*, 48(3):2100–2113.

[7] Del Ser J, Gil-Lopez S, Perez-Bellido A, Salcedo-Sanz S, Portilla-Figueras JA. [2011] IEEE 73rd Vehicular Technology Conference (VTC Spring). On the application of a novel hybrid harmony search algorithm to the radar polyphase code design problem (IEEE Computer SocietyBudapest, Hungary. pp. 1–5.

[8] Gil-Lopez S, Ser JD, Salcedo-Sanz S, Perez-Bellido AM, Cabero JM andPortilla-Figueras JA.[2012] A hybrid harmony search algorithm for the spread spectrum radarpolyphase codes design problem, *Expert SystemApplication* 39(12):11089–11093

[9] Perez-Bellido AM, Salcedo-Sanz S, Ortiz-Garcia EG, Portilla-Figueras JA, Lopez-Ferreras F.[2008] A comparison of memetic algorithms for the spread spectrum radar polyphase codes design problem, Engineering Applications of Artificial Intelligence.21(8):1233–1238.

[10] Karaboga D, Basturk B, [2008] On the performance of artificial bee colony (ABC) algorithm. *Applied soft computing*, 8(1):687–697.

[11] Diwold K, Aderhold A, Scheidler A andMiddendorf M.[2011] Performance evaluation of artificial bee colony optimization and new selection schemes. *Memetic Computing*. 3(3):149–162.

[12] Zhang X, Zhang X, Ho SL and Fu WN.[2014] A modification of artificial bee colony algorithm applied to loudspeaker design problem. *IEEE Transactions on Magnetics*, 50(2): 737–740.

[13] Karaboga D andGorkemli B.[2014] A quick artificial bee colony (QABC) algorithm and its performance on optimization problems. *Applied Soft Computing*. 23:227–238.

[14] Zhang X, Zhang X, Yuen SY, Ho SL, Fu WN.[2013] An improved artificial bee colony algorithm for optimal design of electromagnetic devices. *IEEE Transactions on Magnetics* .49(8):4811–4816.

[15] Zhang X, Wu Z.[2015] Advances in Swarm and Computational Intelligence. Lecture Notes in Computer Science, 9140, ed. by Y Tan, Y Shi, F Buarque, A Gelbukh, S Das, and A Engelbrecht. An artificial bee colony algorithm with history-driven scout bees phase. pp. 239–246.

[16] Karaboga D, Gorkemli B, Ozturk C and Karaboga N.[2014] A comprehensive survey: artificial bee colony (ABC) algorithm and applications. Artificial Intelligence Review 42(1):21–57.

[17] Simon D. [2008] Biogeography-based optimization. *IEEE transactions on evolutionary computation*, 12(6):702–713

[18] Schwefel HP. [1981] Numerical Optimization of Computer Models (Wiley, Chichester).

[19] Dukic, ML Dobrosavljevic, ZS. [1990] A method of a spread-spectrum radar polyphase code design. *IEEE Journal on Selected Areas in Communications*. 8(5):743-749.